Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Environmental Sciences 10 (2011) 1998 – 2004

Procedia Environmental Sciences

2011 3rd International Conference on Environmental Science and Information Application Technology (ESIAT 2011)

Research on Maoershan Stand Factors and 3s Information Based on Ridge Regression Model

Qiang Wang*, Haiqing Hu and Yang Zhang

Northeast Forestry University, No.26 Hexing Road Xiangfang District, Harbin, P.R.China 150040

qiang23709@163.com

Abstract

Taking Maoershan experimental forest of Northeast Forestry University as the study area, and based on the ridge regression analysis as well as a small amount of field survey data and its corresponding remote sensing & GIS information, this paper has screened and optimized the remote sensing and GIS factors affecting the estimation of stand factors such as mean DBH and biomass, found out TM (4×3)/7, TM4/3 and altitude, etc. as the main estimation-affecting factors, established a regression estimation model of stand factors such as mean DBH and biomass by taking pixel as the basic element, and validated the accuracy of the model, thereby proving that the ridge regression estimation model can be used for quantitative estimation of regional forest stand factors.

Keywords: stand factors; remote sensing; ridge trace analysis.

1. Introduction

1.1 Background

The study on stand factors could lay a foundation for studying forest fuels, so the quantitative research on forest stand factors is the fundamental basis and the most important task for modern forest management. Research data can be indirectly used for studies on managing surface fuels and forecasting forest fire, forest fire behavior and forest fire danger. Besides, the data can be further studied for indirectly estimating the stand factors. Some scholars once established an estimation model for the foregoing purpose by adopting multiple regression method [1-2], which is a method that takes the load determined in ground surveys as dependent variables, takes biomass and mean DBH and other stand description factors as independent variables, and at the same time takes into account the information such as slope degree, slope position and the group of dominant tree species. With the continuous development of remote sensing technology, there has been a substantial increase in remote sensing resolution and a constant expansion in spectral range, therefore providing a guideline to estimate forest stand factors in

doi:10.1016/j.proenv.2011.09.313

large space and scale. Foreign scholars have used remote sensing and GIS data to estimate stand factors such as mean DBH through mathematical methods [3-4].

Although the multiple regression model can well estimate the accuracy, multicollinearity would inevitably exist in multiple regression models, so the key point to adopt remote sensing for estimating forest stand factors lies in how to reasonably select major variables, how to overcome the multicollinearity that might exist in selected variables, and how to establish a stable and reliable equation for estimating stand factors. This paper has made an exploration on whether ridge regression can be used to determine variables and then estimate the stand factors such as biomass and mean DBH, and studied the feasibility of applying the ridge regression in the research on stand factors such as biomass and mean DBH.

1.2 Overview on the Study Area

The study area is located in Maoershan experimental forest of Northeast Forestry University, with an geographical coordinate of east longitude 127° 30' - 127° 34' and north latitude 45° 20' - 45° 25'. It has an average elevation of 300m gradually rising from south to north, a general slope of 10 - 15°, and a total forest area of 26507hm2. This area belongs to temperate monsoon climate with distinct seasons of long winter and short summer. Concentrated rainfall occurs in July and August with an annual average rainfall of 723.8mm, annual evaporation of 1094mm, annual average humidity of 70%, annual sunshine hours of 2471.3h, and frost-free period of 120 - 140d.

The area is divided into ten working circles, and its vegetation belongs to Changbai flora, which is a typical kind of natural secondary forest in northeast and eastern mountain areas. The area originally possessed a zonal climax community of Pinus koraiensis broad-leaved forest, which has been destructed and retrograded to the current natural secondary forest. It holds a forest coverage rate of 70.2% with diverse range of forest types and various woody and herbaceous plants of more than a thousand species. The main arbor species include Larix olgensis, Betula platyphylla, Quercus mongolica, Populus davidiana, Tilia amurense, Pinus koraiensis, Juglans mandshurica, Fraxinus mandshurica, etc.. The region’s zonal soil is dark brown forest soil, and in areas with perennial or seasonal pooling water, there is dark brown forest soil undergone the process of gleization and the affection of white slurry and meadows.

2. Study Methods

2.1 Field Investigation and Laboratory Test

We have carried out field survey, established sample plots, selected representative forest types, and totally set 40 sample plots sized 20 × 20m2 through the method of mechanical distribution. Conventional survey methods were applied when measuring the stand and site factors such as slope, aspect, canopy density, DBH, tree height, age, etc., and the data of totally 179 sample plots of continuous forest inventory were collected in Maoershan Forest Farm within the region covered by TM images in 1990.

Mean DBH, average tree height and average age were calculated and recorded.

2.2 Data Processing

Data processing included the processing of remote sensing data and the processing of GIS data. The U.S. Landsat-7 land resources satellite ETM+ data were used as remote sensing data, and its track number is 117/28. Through bilinear interpolation, it used ERDAS8.6 remote-sensing image-processing software to precisely and geometrically correct the TM images, and had 220 control points in the images. There was an error of 0.2650 pixels in the longitude direction, an error of 0.2992 pixels in latitude, and an error of 0.3997 pixels in total.

Based on the geographic coordinates, gray values were obtained from the grayscale binary data files corresponding to corrected images.Suppose the geographical coordinate of starting pixel on the post-corrected remote sensing image as (x0, y0), the geographical coordinate of sample-plot information as (x1, y1), and the followings are satisfied:

$$ABS(x_1 - x_0) \le 15 \tag{1}$$

$$ABS(y_1 - y_0) \le 15 \tag{2}$$

Through VB programming, we could read the gray values of pixels corresponding to all fixed sample plots. At the same time, we could read the following four positional gray values: $(x_0 - 30, y_0)$; $(x_0 + 30, y_0)$; $(x_0, y_0 + 30)$ and $(x_0, y_0 - 30)$, and acquire the mean value with the gray value of (x_0, y_0).

By using the DEM data obtained from Maoershan experimental forest and combining the coordinates established in field investigation, GIS data could lead to GIS information such as the slope, altitude of pixels to corresponding sample plots.

2.3 The Setting of Independent Model Variables

Independent variables are the information of ground sample plots obtained by RS and GIS, including: the gray value of each band, gray ratio, and the sample plots' vertical & horizontal coordinates, altitude, land type, slope, aspect, canopy density, etc., while dependent variables are the measured stand factors on ground sample plots. Setting dependencies can be summarized as follows based on:

1 Selecting the gray values as independent variables

Gray values of each band on TM images are closely related to the spectral reflectance properties of vegetation, vegetation density, vegetation growth and other factors of soil conditions, so TM1, TM2, TM3, TM4, TM5, TM7 were selected as independent variables that may affect the estimation of stand factors.

2 Selecting the ratios of gray values and vegetation index as independent variables

Considering that the characteristics of TM remote sensing data in each band and the spectral reflectance properties of vegetation show a linear correlation with the density of plant distribution, so ratios of gray values and vegetation indices such as TM7 / 3, TM (5+7-2) / (5+7 +2), TM (4 × 3) / 7, TM3 / Σ could be set as variables.

Taking into account that vegetation indices are dominated by combinations of red-band and near-infrared band, involving more than 90% of the vegetation information, the vegetation index can be quantitatively measured to indicate vegetation vigor, and it has better sensitivity and anti-interference than single band when being used to detect the biomass. The following independent variables were adopted: vegetation index of normalized difference NDVI = TM (4-3) / TM (4 +3), ratio vegetation index RVI = TM4 / 3 and environmental vegetation index EVI = TM (4-3).

3 Selecting GIS factors as independent variables

The size of stand factors has some correlation with the location, so GIS factors such as altitude, slope and aspect are the factors for estimating stand factors.

2.4 Model Establishment

1 Establishment of ridge regression model

Since a large amount of impact factors for estimating forest stand factors were selected as independent variables, approximate linear relationship may exist among these independent variables, such as band ratios, and it's quite possible that the universally-applied least squares estimation cannot produce optimal solution. Therefore, it's essential to overcome these

Table 1. Variables to be selected

$x_{1\sim6}$	Variables	$x_{5\sim10}$	Variables	$x_{11\sim15}$	Variables	$x_{16\sim20}$	Variables
x_1	Vertical Co.	x_6	TM4	x_{11}	Altitude	x_{16}	TM(5+7-2)/(5+7+2)
x_2	Horizontal Co.	x_7	TM5	x_{12}	TM4/3	x_{17}	TM(4+5-2)/(4+5+2)
x_3	TM1	x_8	TM7	x_{13}	TM(4×3)/7	x_{18}	TM(4×5)/7
x_4	TM2	x_9	Slope	x_{14}	TM(4-3)/(4+3)	x_{19}	TM3/Σ
x_5	TM3	x_{10}	Aspect	x_{15}	TM7/3	x_{20}	Canopy Density

possible approximate linear relationships in order to get accurate model for estimating stand factors. To solve this problem, ridge regression model was established to estimate the stand factors, and observation matrix X could be made up according to the established factors dominating the estimation of forest stand factors and the values of main factors to corresponding sample plots. To facilitate regional estimation, it's unnecessary to centralize and standardize the candidate factors when establishing the actual estimation equation. The Ridge regression estimation model linking stand factors and all factors are as follows:

$$Y = X\gamma(k) + e \tag{3}$$

In formula (3), Y represents the observation vector composed by the measured stand factors of sample plots; X_1 represents the observation matrix of impact factors, $\gamma(k)$ represents ridge regression coefficients, and e represents the observation error of vegetation coverage in sample plots. Ridge parameter k was solved according to the ridge trace map, and based on the principle of ridge regression, the equation to solve ridge regression values of main impact factors could be identified as follows:

$$\gamma(k) = (X^T X + kI)^{-1} X^T Y \tag{4}$$

Formula (4) can lead to the undetermined coefficients of estimation factors, and then to the ridge regression estimation model of stand factors.

2 Selection of variables in ridge regression model

Ridge trace map can directly reflect both the impact of variables on the estimation of forest stand factors and their mutual relationship. Through the ridge trace map and characteristic analysis, we can effectively select the major remote sensing & GIS information affecting the impact of forest stand factors. Independent variables selected by ridge trace should follow the following principles:

(1) Removing the independent variables with both stable ridge regression coefficient and small absolute value. When calculating the ridge regression values of undetermined parameters, the observation matrix X had been centralized and standardized. The ridge regression coefficients of variables can be directly compared.

(2) Removing the independent variables with unstable ridge regression coefficient but rapidly tending to 0 with the increase of ridge parameter k.

(3) Based on the characteristic roots of square matrix $X^T X$, possible approximate linear relationship among the set variables impacting the estimation of forest stand factors can be analyzed according to the characteristic roots close to 0. By using the approximate linear relationship and combining with the ridge trace of variables contained in such relationship, we removed one or more independent variables with unstable ridge regression coefficients. Except the independent variables with small ridge regression coefficients, generally, the number of independent variables that should be removed equals to the number of approximate linear relationships.

3 Prediction of model accuracy

In this study, absolute values of average relative errors were applied for the evaluation on model, and the formula [5] is as follows:

$$M_{APE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{5}$$

3. Results and Analysis

3.1 *Model Results for Biomass Estimation*

1 Selected results of independent variables

According to the above methods for selecting variables, we read the values of 20 factors being listed by 31 sample plots contained in the remote sensing image for ridge trace analysis, and the ridge trace of 20 factors shown in Figure 1 can be plotted by taking forest biomass as dependent variables.

As can be seen from the ridge trace map of Figure 1, the independent variables of altitude and canopy density have relatively stable ridge regression coefficients with small absolute values, so these variables can be removed according to the first principle of variable selection. The variables of TM1 and TM (5+7-2) / (5+7 +2) have Unstable ridge regression coefficients and would verge to zero quickly with the increase of k, so these variables should be removed according to the second principle.

So at last, 12 variables including TM2, TM3, TM4, TM5, TM7, TM4/3, TM(4×3)/7, TM(4-3)/(4+3) , TM3/Σ,aspect,slope,altitudewere left as the main factors impacting the estimation of forest biomass. The setting of independent variable is shown in Table 2.

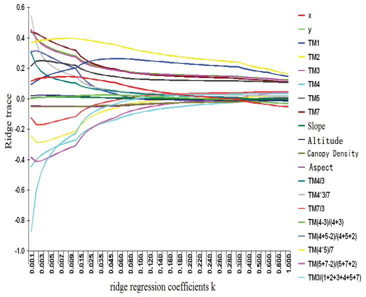

Fig.1. Ridge trace map of candidate factors

Table 2. Candidate model independent variables

$x_{1\sim6}$	Variables	$x_{7\sim12}$	Variables
x_1	TM2	x_7	Aspect
x_2	TM3	x_8	Altitude
x_3	TM4	x_9	TM4/3
x_4	TM5	x_{10}	TM(4×3)/7
x_5	TM7	x_{11}	TM(4-3)/(4+3)
x_6	Slope	x_{12}	TM3/Σ

2 Model fitting

As can be seen from the ridge trace map plotted above, when the ridge parameter $k = 0.5$, the ridge trace map of candidate factors was basically stable. By determining the solution equation of major impact factors according to ridge estimation principles, we could obtain the undetermined coefficients of 12 estimation factors and then get the model for biomass estimation. The error of biomass estimation model is shown in Figure 2: an accuracy of around 89%.

a) *Results From the Model of Mean DBH Estimation*

According to the above estimation methods, we read the values of 20 factors being listed by 31 sample plots contained in the remote sensing image. As shown in Table 1, the square matrix $X^T X$ was calculated from the observed matrix X after the centralization and standalization. The square matrix $X^T X$ was used to calculate the characteristic roots of 19 factors and the vectors of corresponding characteristic roots.

1 Selected results of independent variables

The ridge traces of the 19 factors shown in Figure 3 could be plotted by taking stand mean DBH as dependent variables. In accordance with the analysis on the ridge trace map, the major remote sensing and qualitative factors impacting the estimation of mean DBH are as follows: TM2, TM4, etc..

As can be seen from the ridge trace map of Figure 3, the independent variables of TM1 and TM (4×5)/7 have relatively stable ridge regression coefficients with small absolute values, so these variables can be removed according to the first principle of variable selection. The variable of TM3 has unstable ridge regression coefficient and it would verge to zero quickly with the increase of k, so these variables should be removed according to the second principle. Similarly, the independent variables of vertical axis, horizontal axis, altitude and aspect can be removed because of their relatively stable ridge regression coefficients with small absolute values.

So at last, 8 variables including TM2, TM4, TM5, TM7, TM7/3, TM(4-3)/(4+3) , TM(4×3)/7, TM3/Σ were left as the main factors impacting the estimation of mean DBH.

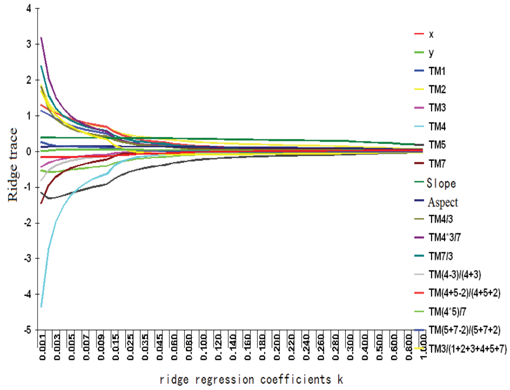

Fig.3. Ridge trace map of candidate factors

Table 3. Independent variables of mean DBH estimation model

$x_{1\sim4}$	Variables	$x_{5\sim8}$	Variables
x_1	TM2	x_5	TM(4×3)/7
x_2	TM4	x_6	TM7/3
x_3	TM5	x_7	TM(4-3)/(4+3)
x_4	TM7	x_8	TM3/Σ

2 Model fitting

As can be seen from the ridge trace map plotted above, when the ridge parameter $k = 0.035$, the ridge trace map of candidate factors was basically stable. By determining the solution equation of major impact factors according to ridge estimation principles, we could obtain the undetermined coefficients of 8 estimation factors and then get the model for estimating stand mean DBH:

Where: y represents the stand mean DBH (cm) with a deviation of around 11%.

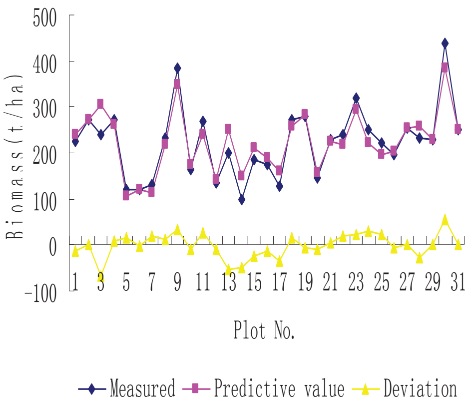

Fig. 2. Comparison of the results from biomass estimation model

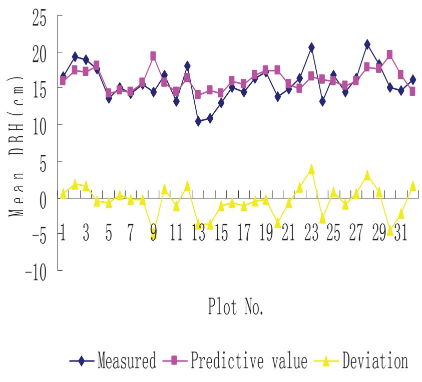

Fig. 4. Comparison of the results from mean DBH estimation model

Conclusions

Through ridge regression model, the remote sensing and GIS data can well estimate the biomass and mean DBH and other stand factors with a good fitting result and fine accuracy, which demonstrates that it is feasible to estimate the stand biomass and mean DBH in certain region by using remote sensing data through the ridge regression model. The estimation accuracy of ridge regression model established by using matlab is higher than that of multiple regression model. Its error distribution curve is smoother than that of regression equation, and has overcome the problem of possible multicollinearity.

Through the ridge trace analysis, the contribution of remote sensing & GIS information which affects the estimation of stand factors to the estimation of forest stand factors as well as their mutual interactions can be quantitatively and visually displayed. Experimental studies show that, when using the remote sensing & GIS information corresponding to a small amount of sample plots to estimate the mean DBH and other stand factors, remote sensing information provides the foundation, and GIS information also plays a very important role, but only by combining two of them, can we make an effective estimation.

When selecting the independent variables of ridge regression, it should be considered that over 90% of canopy density of sample plots mainly distributes between 0.4-0.7, so the regression model would lead to considerable error if being used for forests with a canopy density of less than 0.4 or greater than 0.7. In addition, due to the variety of stand forest types, regional environmental factors and biological factors, the estimation model can only be modified and applied to this region.

The information obtained from remote sensing materials basically comes from ground radiation. When the canopy density of stand is greater than 0.8, most effective radiation of undergrowth grass will be scattered and reflected by the canopy, resulting in estimation bias, so it is predicted that the correlation between remote sensing information and undergrowth stand factors will be diminished at this time.

References

[1] Liu Xiaodng, Wang Jun,and Zhang Dongsheng: Forest fire. Vol.3 (1995),p.8

[2] Zhao Xianwen, in: Quantitative Estimation of Forest Resource Based on “3S”, edtied by China Science and Technology Press, Beijing (2001), in press.

$$y = 2.9920 + 0.0175x_1 - 0.0677x_2 - 0.3372x_3 + 0.5945x_4 + 0.1622x_5 + 3.0375x_6$$
$$-0.9977x_7 + 0.5778x_8$$

[3] Nelson R, Latty R, Mott G.:Photo gram metric Engineering and Remote Sensing. Vol.50 (2002),p.607-617

[4] E.A Catchpole:International Journal of Wildland Fire.Vol.10(2) (2001),p.215-222

[5] John W. Modeling Potential Erosion due to the Cerro Grande Fire with a GIS-based Implementation of the Revised Universal Soil Loss Equation. International Journal of Wildland Fire.Vol.12(1) (2000),p.85

[6] Henry Lachowsk. A Primer on Mapping Vegetation Using Remote Sensing. International Journal of Wildland Fire.Vol.10(4) (2002),p.277